DOCUMENT RESUME

ED 449 205 TM 032 315

AUTHOR Bracey, Gerald W.
TITLE A Short Guide to Standardized Testing. Fastback 459.
INSTITUTION Phi Delta Kappa Educational Foundation, Bloomington, IN.
ISBN ISBN-0-87367-659-9
PUB DATE 2000-00-00
NOTE 57p.; Sponsored by the Northern Virginia Chapter of Phi Delta Kappa International.
AVAILABLE FROM Phi Delta Kappa International, P.O. Box 789, Bloomington, IN 47402-0789. Tel: 800-766-1156 (Toll Free); Fax: 812-339-0018; Web site: http://www.pdkintl.org.
PUB TYPE Guides - Non-Classroom (055) -- Reports - Descriptive (141)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Achievement Tests; Elementary Secondary Education; *Standardized Tests; *Test Construction; Test Interpretation; Test Use

ABSTRACT

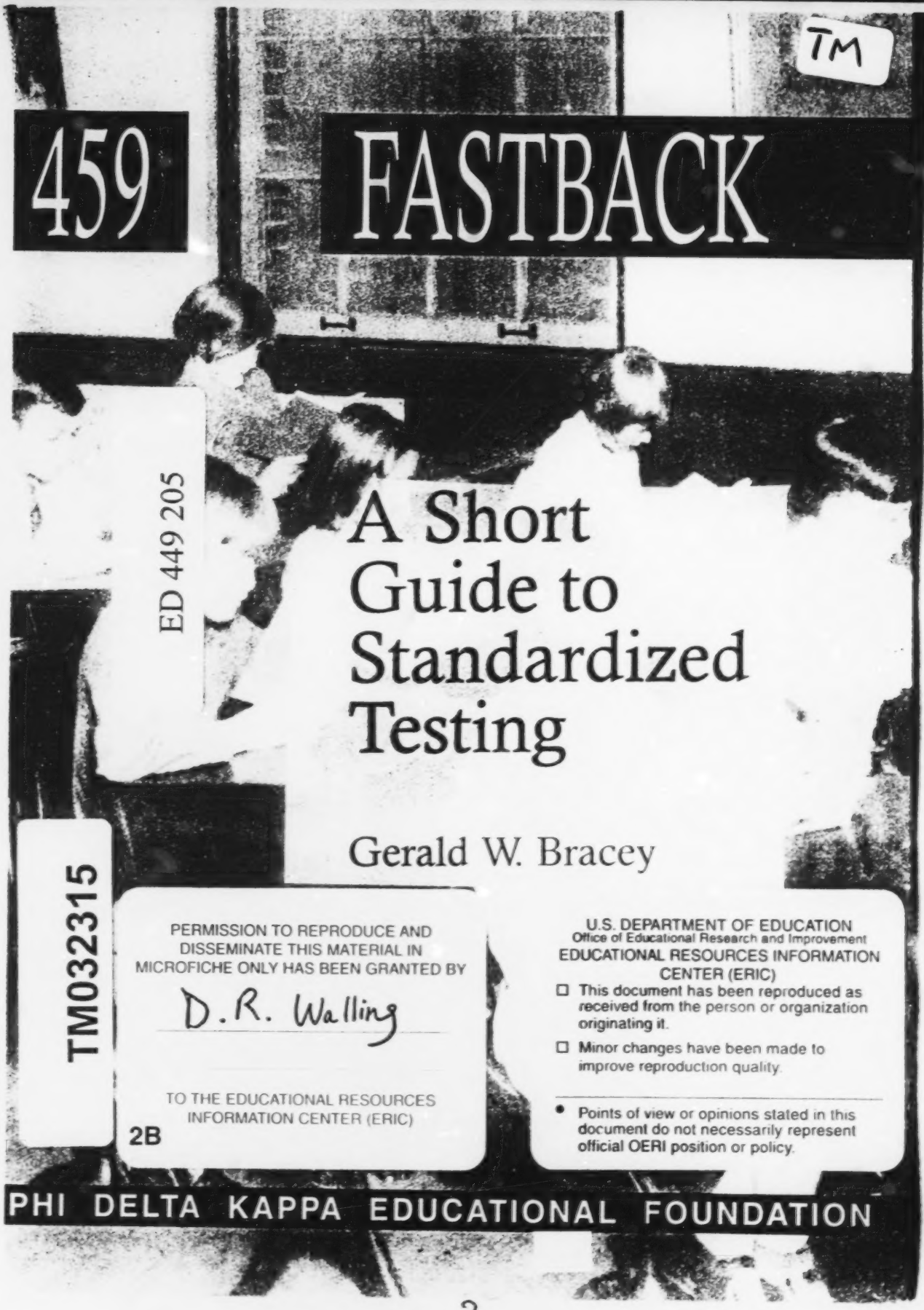
This fastback provides information about what tests can and cannot do, how they are constructed, and how they are used and misused. It offers suggestions about how best to interpret test results. The chapters are: (1) "A Test on Testing"; (2) "Basic Considerations"; (3) "Standardized Tests"; (4) "Performance Tests"; (5) "Interpreting Test Scores"; and (6) "Conclusion." (SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

TM

459

FASTBACK



ED 449 205

A Short Guide to Standardized Testing

Gerald W. Bracey

TM032315

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

D.R. Walling

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PHI DELTA KAPPA EDUCATIONAL FOUNDATION

Gerald W. Bracey

Gerald W. Bracey is a well-known policy analyst, researcher, and writer, focusing primarily on testing and assessment. He writes a monthly column on educational and psychological research for the *Phi Delta Kappan*. Since 1991 he has produced reports on the condition of the public schools that also appear in the *Kappan*. The first seven of these reports are collected in a volume titled *The Truth About America's Schools: The Bracey Reports, 1991-97* (Phi Delta Kappa Educational Foundation, 1997).

Early in his career Bracey was a research psychologist in the Early Childhood Education Research Group at the Educational Testing Service. He then became associate director of the Institute for Child Study at Indiana University. Later he became director of Research, Evaluation, and Testing for the Virginia Department of Education and, after nine years, moved to a similar position with the Cherry Creek school district near Denver, Colorado.

Bracey holds a Ph.D. in psychology from Stanford University. This fastback is condensed from his book, *Put to the Test: An Educator's and Consumer's Guide to Standardized Testing* (Phi Delta Kappa International, 1998).

Series Editor, Donovan R. Walling

A Short Guide to Standardized Testing

by
Gerald W. Bracey

ISBN 0-87367-659-9
Copyright © 2000 by the Phi Delta Kappa Educational Foundation
Bloomington, Indiana

This fastback is sponsored by the Northern Virginia Chapter of Phi Delta Kappa International, which made a generous contribution toward publication costs.

Table of Contents

Introduction 7

A Test on Testing 9
The Answers 10

Basic Considerations 14
The Uses of Standardized Testing 15
Secrecy 19
Factors Influencing Test Scores 20
Aptitude, Ability, and Achievement Tests 23

Standardized Tests 26
How Standardized Tests Are Constructed 27
Norm-Referenced Tests 28
Criterion-Referenced Tests 35

Performance Tests 39

Interpreting Test Scores 45
Grade-Equivalents 45
Percentiles 47
Normal-Curve Equivalents 48
Standard Scores 48
Stanines 51

Conclusion 52

Notes 53

Introduction

Tests are everywhere. Why have they become so important? The short answer is that people lost confidence in their schools and the people who run them. For those worried about the quality of the schools, the question became, If we can't trust people in schools to tell us about how well they are functioning, what can we trust? In looking around for means by which to evaluate schools, various outsiders discovered tests.

Tests were external. Tests seemed objective. Best of all, tests in their multiple-choice formats were cheap. And once electronic scoring of answer sheets became a reality, the results could be known quickly. Suddenly, there were tests for basic skills, for high school graduation, for teacher certification, and for accountability.

Unfortunately, the increased use and reporting of tests has not always been accompanied by increased understanding of how tests can or ought to be used. Teachers need to know what the tests can and cannot say about children. They need to be able to evaluate whether the decisions made about children (and themselves) on the basis of test scores are fair and appropriate. Administrators need to know about tests to determine

what policy decisions about the district — and the administrators in it — are appropriate.

There is, then, a need for clear and objective information about what tests can and cannot do, about how they are constructed, and about how they are used and misused. It is my hope that this fastback will prove to be a source of such information.

A Test on Testing

Take the following true-or-false quiz. The answer and a short explanation for each question follow the test, and more details can be found throughout this fastback.

1. If a school district's high school graduates do not all read at grade level, the district is not doing its job of educating the students.
2. If the average total SAT score at the university you most want to go to is 1100 and your total SAT score is only 1040, you shouldn't bother to apply.
3. If in the Mark Twain School, Mrs. Smith's students do not score as high as Mrs. Jones' students, Mrs. Smith is not as good a teacher as Mrs. Jones.
4. If a school adopts a new curriculum and test scores fall, the new curriculum is shown to be inferior to the old curriculum.
5. The scores on a test naturally fall along a "normal"— bell-shaped — curve.
6. If two curricula are compared and the students using curriculum A score statistically significantly higher than the students using curriculum B, then curriculum A is the better curriculum.

7. A fourth-grade student reading at seventh-grade level should be promoted to the seventh grade, at least for reading instruction.
8. The SAT is a "common yardstick," meaning that two students with the same score from different states have the same potential for college.
9. If a child who scored at the 63rd percentile in reading one year falls to the 57th percentile in the next year, something is wrong. Either the child isn't trying, or the second year's teacher is not as effective.
10. Research indicates that IQ is 80% inherited.

The Answers

All of the statements in the quiz are false.

1. Most tests define "grade level" as the score of the average child at a given grade. For a particular school, all children could possibly be above grade level — if the school were excellent or located in an affluent neighborhood. But nationally half of all students are, by definition, below average — that is, below grade level.

2. If the average total SAT score is 1100, this means that half the incoming freshmen are accepted with scores below this average.[1] Your 1040 might well be acceptable. In addition, you are not in competition with all other applicants. If that were true, universities would not only be unable to field athletic teams, but they also would be unable to have fine arts or performing arts departments. Like athletes, students with these special talents often have trouble with paper-and-pencil tests. However, if

10

your parents went to the same school or can afford to pay all of your expenses, your chances of admission are mightily improved.

3. There are many possible reasons why the classes could differ. Mrs. Smith might not emphasize material that the test covers as much as Mrs. Jones does. Mrs. Smith also might be a newer teacher; veteran teachers often are rewarded by being assigned to high-achieving classes.

4. The new curriculum might not match the test as well as the old curriculum did. Tests measure highly specific aspects of a curriculum in highly specific ways. One study found that, three years after a district changed tests, the students did not score nearly as well on the old test as they did when it was used routinely.

5. Many tests will fall along a bell-shaped curve, but often such a curve is forced on the test data by the test maker through the use of "item-selection techniques" and other statistical procedures. Many educators strive to make scores fall along a curve that looks like the one in Figure 1 on the next page, often called a "j" curve because it resembles the letter "j." On a bell curve, most students have learned something, a few have learned a lot, and a few have learned little. On a "j" curve, most students have learned a lot, and only a few have learned little.

6. A statistically significant difference might have no practical implications at all. Remember that a statement of statistical significance is a statement of odds: how likely it is that the difference between curriculum A and curriculum B would be seen if the two curricula were really the same. (If this seems puzzling now, not to wor-

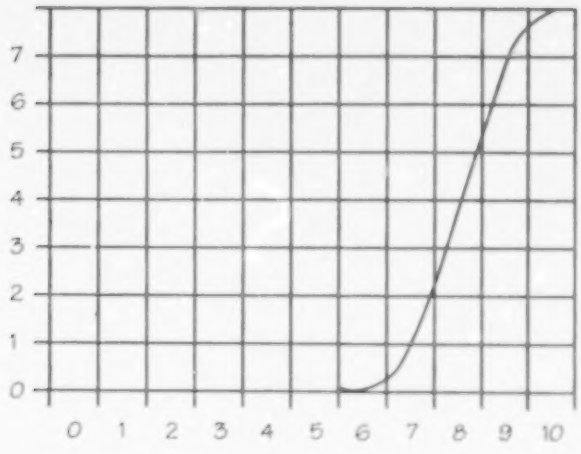


Figure 1. A J curve.

ry; this is the most technical concept in the book and will be explained in some detail.)

7. The fourth-grade student with a seventh-grade reading level has the same score as the average seventh-grader would get when reading fourth-grade material, not seventh-grade material.

8. Consider two students. One comes from a well-educated, affluent family and attends an elite, private, college-preparatory high school. This student scores 600 on the SAT verbal. The other student comes from an impoverished inner-city neighborhood, attends a public school with few books, has no quiet place to study, and works to help the family make ends meet. This second student also scores 600 on the SAT verbal. Are these two students equally likely to be successful in college?

9. I don't think any published data exist to answer this question directly; but when I looked at the scores of elementary school students over a four-year period, most students varied over that period by about 25 percentile ranks. Those above the 95th percentile varied

less, but only those at the 99th percentile and those below the 10th percentile tended to have really stable scores.

10. Researchers differ wildly over the relative influence on IQ of genetics and environment. What's more, the whole question might be meaningless.

Basic Considerations

Tests in this country have historically been used to sort people: officers and enlisted men, college material and vocational school material, gifted and talented and the rest of us, bluebirds and robins. In fact, the schools themselves have sometimes been called "The Great Sorting Machine." Some people argue that sorting is a fundamental purpose of schools — to get people into the appropriate societal slots. Others argue that the schools should educate all children to the highest possible levels. Both positions have had prominent followers, and the debates between the two positions have been heated, to say the least.

The most obvious use of testing as a sorting device occurs in tracking, wherein some students are permitted to study more advanced topics than are other students. The early testers saw tracking as humane: to confront a child of low ability with the same curriculum as provided to a child of high ability would only frustrate and humiliate the low-ability student. This attitude exists in many places today, though its expression is muted because in the tenor of these times it is not "politically correct." And, in fact, it might also be wrong.

The Uses of Standardized Testing

This section discusses the common uses of standardized tests, including those uses that are really not legitimate.

Monitoring. Tests are sometimes used by teachers and parents as a kind of "reality check" to see if the test results for a child accord with other indicators of achievement, such as classroom performance and report card grades. Testers like to emphasize that tests are external to the classroom and "objective," in contrast to the "subjective" judgments of teachers. A teacher's judgments are fallible, of course, but so are tests; and the teacher's "subjective" judgments are based on her observing the child over a much longer period of time. Thus the tests can be misused in this fashion if people put too much faith in the test over the other information.

Diagnosis. While there are tests designed to diagnose particular conditions that might merit special education programs for a child, it is very difficult to use the typical commercial achievement test in a diagnostic fashion. There are too few items in any one skill area to give a very reliable indication at the level of the individual child. It is difficult to assess particular problems in reading comprehension, for example. In the area of arithmetic, there is more opportunity for diagnostic work because the skills are more precisely defined. It is possible to see, for instance, if a child is having difficulty with place value or with converting fractions to decimals. Even here, though, the test is not very diagnostic and certainly not prescriptive. The best analogy

15

might be to liken the test to a thermometer. An above normal temperature tells you that something is wrong, but, by itself, it provides no clues about what is wrong and suggests no prescription for what to do about it.

Teacher Accountability. This is one of the most seductive uses of tests because, on the surface, it seems so reasonable. Of course a teacher should be accountable for what his or her students know. However, there are a variety of problems with using tests for teacher accountability. In the first place, one teacher might emphasize what is on the test while another prefers to teach other subject matter and can provide good justifications for that offering. Not everything can be taught, and two teachers may well differ on how to teach the same material. One could elect to teach science through physics, another through ecology. One could elect to present science as a series of disciplines, another could emphasize science as a process and focus on the scientific method that cuts across disciplines. Of course, if the tests are used for accountability, teachers will come to emphasize what is on them.

There are other problems with tests and teacher accountability. The process collides with the common system of assigning teachers to classrooms. Logically, one might think that the best teachers would be assigned the toughest classes. In many places, the reverse holds. Furthermore, children do not arrive in a teacher's classroom as blank slates. They come with previous levels of achievement. These levels of achievement greatly depend on what happens and has happened outside of school. From birth to age 18, an American

child is in school 9% of his or her life. The external environment affects test scores a great deal.

Principal/Superintendent/Board Accountability. There are states, such as Michigan and North Carolina, where the consequences of low test scores are visited not on the teachers but on one or more of the school's or district's administrators. These programs are too new to know what effect they will have. In North Carolina, if a school is declared low achieving, it receives assistance from the state and its principal is put on probation. In one instance where this happened in 1997, the district sued the state. All the considerations mentioned under teacher accountability apply here as well.

Student Accountability: Promotion, Retention, and Graduation Decisions. There appears to be a growing trend in this country to use tests for decisions about promotion and retention and about eligibility for graduation. About retention in grade, this much can be said: It doesn't work. Study after study has found the consequences to be negative. One study ranked 49 education innovations in terms of their effect on achievement. Retention in grade ranked 49th. It was among the few innovations that actually produced negative results.

Tests currently used for graduation eligibility are for the most part not commercial achievement tests, but tests constructed ostensibly around state curricula. Students typically get six or more attempts at them before graduation time so the likelihood of a student not passing despite having an achievement level well above what the test reports is quite small. Whether the test is testing what high school students should know is an-

other question, one that needs to be evaluated in each instance.

Selection Decisions. As noted at the beginning of this chapter, tests were designed to make discriminations among people in order to provide some people one set of educational experiences (officer candidate school, college entry, a more challenging curriculum, etc.) and other people different educational experiences. The hope has always been to match the experiences to people's needs and abilities, but it has not always worked out that way.

In selection for college, admissions officers have a great deal of information about students, often going beyond SAT (or ACT) scores, grades, and rank in class. Many colleges now ask for essays or personal interviews with alumni, and students themselves sometimes send in videotapes. Actually, as many observers have pointed out, there is little evidence that most colleges are "selecting" students. From 1977 to 1994, the number of high school graduates declined each year as the baby boom bottomed out. Yet college enrollments rose by more than four million during the same period.

In selecting students for gifted and talented programs or other enrichment programs, tests often play a dominant role — sometimes the only role. Selection into special education programs is not usually quite so test dependent. This is good because the tests used for diagnosis of special needs, though they are specialized tests, often do not have as high reliability as the typical achievement tests. Decisions for special education selection are made, or at least should be made, in a group

consultation involving teachers, the parents, and the special education specialists of the district.

Secrecy

If you are surprised to see a segment on secrecy (it's usually called "test security") because you don't think it's an issue, then your next reaction should be fear because you have been seduced into accepting as natural a most unusual, even pathological situation. In *Assessing Student Performance*, Grant Wiggins captured the setting and our all-too-usual casual reaction to it well:

> It is so common that we barely give it a second thought: the tests that we and others design to evaluate the success of student learning invariably depend upon secrecy. Secrecy as to the questions that will be asked. Secrecy as to how the questions will be chosen. Secrecy as to how the results will be scored. Sometimes secrecy as to when we will be tested. Secrecy as to what the scores mean. Secrecy as to how the results will be used. What a paradoxical affair! Our aim is to educate, to prepare, to enlighten, yet our habits of testing are built upon procedures that continually keep students in the dark — procedures with roots in premodern traditions of legal proceedings and religious inquisitions.[2]

Actually, the situation is worse than that because Wiggins has only the students in mind. But most of the time the teachers also do not know what is going on.

Wiggins notes that the secrecy aspect of testing helps produce people who are both docile (they put up with it) and leery. The point also is made by anthropologist

F. Allan Hanson in his book, *Testing Testing: Social Consequences of the Examined Life*:

> Whether the results are positive or negative is irrelevant. The point is that testing opens the self to scrutiny and investigation in ways that the self is powerless to control. So far as the areas of knowledge covered by the test are concerned, this transforms the person from autonomous subject to passive object.[3]

Factors Influencing Test Scores

Discussions about tests in the media often make it seem like only what happens in school has any effect on test scores. This is hardly true. After all, a child spends only 9% of his or her life from birth to age 18 in school. We should keep in mind the following factors in thinking about test scores.

Family Income. If you look at the College Board's Profiles of College-Bound Seniors, which comes out each fall with the release of SAT scores, you will see one table showing SAT scores by income level. There is a very clear progression: the higher the income, the higher the SAT score.

Educational Level of Parents. In many studies, the parents' educational level is the single biggest factor contributing to the test scores of children.

Poverty. One study examined scores in high- and low-poverty schools. High-poverty schools were defined as those in which at least 76% of the students were eligible for free or reduced-price lunches. Low-poverty schools had zero to 20% eligibility rates. The researchers

first divided the students into groups depending on the kinds of letter grades they took home on report cards. Then they looked to see how these groups performed on standardized tests of reading and math. Students in low-poverty schools who got A's scored about as you would expect, averaging the 81st percentile in reading and the 87th in math. Students in high-poverty schools who got A's scored higher than students who got lower grades, but in neither case did the average score reach even the 40th percentile.

Motivation. Motivation can have an enormous impact on scores. Being motivated to do well on a test is no small matter. To make students more serious about the tests, one superintendent decided to take them out of the academic arena. He presented the tests, not as an opportunity for the students to show how smart they were or how well their teachers had taught them, but as a chance to beat their archrivals in the adjacent county, just as they tried to do each year in sports. During the week of testing, the teachers dressed as cheerleaders and led pep rallies in the auditorium, where the students in the affected grades were cheered on by their nontested peers. The motivational program worked. Depending on age and test topic, the scores were 15 to 30 percentile ranks higher than the previous year's marks.

Personal Hygiene. I use this term to cover such things as getting a good night's sleep before the day of testing and eating well on the day of testing. Hungry children do not score as high as well-fed children do.

Cultural Factors. This is another global category to cover a number of influences. In the Third International

21

Mathematics and Science Study (TIMSS), most of the 41 countries participating, especially Western countries, had eighth-grade test scores that were very similar to one another. Five or six developing nations scored at the bottom, and four Asian nations — Singapore, Japan, Korea, and Hong Kong — scored high. (Taiwan did not take part; it probably would have been another high scorer.) In between were about 30 nations with few differences. In the high-scoring nations it is not unusual for students to go to classes after school or to go to a private tutor. They also go to school on weekends. Here's a typical report from Japan:

> Akiko Tsutui, a 10-year-old fifth-grader, gets out of school at 3:30 p.m. and goes straight home to have a snack and do her homework. Three afternoons a week she leaves again at 4:45 for a juku (cram school for tests) session that lasts from 5:10 to 10:00. For almost the entire class, Akiko will listen to tutors explain how to answer test questions and will practice taking them herself.[4]

Students in these countries also have additional motivations. In Japan, for example, it is critically important for the future that children get into the "right" high school and even more important to get into the "right" college. In the January 1998 issue of *Principal*, Kazuo Ishizaka, president of the Japanese Council on Global Education, writes, "Because Japanese society judges people on the basis of the schools they attended rather than their ability and skills, Japanese parents try to send their children to the best schools."[5]

When I lived in Hong Kong, the exams given at the end of eighth grade were even more life-determining

22

because there was room in high schools for only 25% of the eighth-grade students. Suicide hotlines were set up to handle calls from those most distraught by their failure to obtain a place in a high school.

Aptitude, Ability, and Achievement Tests

Few notions in testing have caused more mischief than the distinctions between aptitude, ability, and achievement tests. Conceptually, the three are indistinguishable. Ability and aptitude tests do not measure something qualitatively different from what achievement tests measure.

The trouble begins when people assume, as they all too often do, that an aptitude or ability test measures "potential." Given a measure of potential, we can then label children as "where they ought to be," "underachievers," or "overachievers," according to how their performance in school compares to their performance on the test. We could with just as much logic label a child an "overtester" or "undertester."

That there is no conceptual distinction between ability tests and achievement tests has been known for many years, but it is something that just doesn't enter into the everyday discussion of tests. When the National Research Council, part of the National Academy of Sciences, conducted an extensive study of ability testing, it found citations as early as 1927 declaring that the two "types" of tests were fundamentally the same. The Council concluded:

> The line of demarcation between aptitude and achievement tests is not as clear-cut as popularly be-

23

lieved. The major difference has been well stated by Anne Anastasi: "Today the difference between these two types of tests is chiefly one of degree of specificity of content and extent to which the test presupposes a designated course of instruction."[6]

The Council's citation from Anastasi, an eminent psychometrician, comes from the 1976 edition of her textbook on testing. More recently, Anastasi affirmed her position.

> We should especially guard against the naive assumption that achievement tests measure the effects of learning while aptitude tests measure "innate capacity" independent of learning. This misconception was fairly prevalent in the early days of psychological testing but has been largely corrected in the subsequent clarification of psychological concepts. It should be obvious that all psychological tests measure the individual's current behavior, which inevitably reflects the influence of prior learning.[7]

An achievement test supposedly looks back in time; an ability or aptitude test supposedly looks forward. In fact, any test can be used for both purposes. We typically use the SAT to predict college success, and we use such tests as the Stanford Achievement Tests to summarize accomplishment in high school. But there is no reason why we couldn't use the Stanford to predict college success. Such prediction is merely a statistical process, the calculation of a correlation coefficient. We take the scores from our test — SAT or Stanford or any other test — and correlate those scores with the freshmen grade-point averages of those we tested.

The real difference between ability and aptitude tests on the one hand and achievement tests on the other is that we have a better idea of how a student came to do well or not so well on an achievement test. Even if an achievement test doesn't wholly reflect the curriculum — for example, a phonics-based reading test given to a whole-language class — it still reflects to some extent what is being learned in school, reading. Ability tests often don't. For instance, the Cognitive Abilities Test (CogAT) includes a "nonverbal" section that requires students to look at a series of geometric forms and to choose from a selection the one that would be next in the series. Some kids are really good at this, and some are awful. But it is unlikely that any of them received instruction in "geometric form prediction" from their teachers.

It is possible that a more meaningful index of ability could be obtained from multiple measures. If we tested a student in the fall and again in the spring and measured how much better she did in the spring than in the fall, we might use this difference as a measure of ability in terms of rate of learning, though it is still not a measure of "potential." Although this idea is plausible, no one has systematically tried to develop this notion of ability. Few people have even dabbled at it.

25

Standardized Tests

Most of the tests that people take or are likely to read about in the newspapers are referred to as "standardized tests." The SAT is probably the best known of such tests. But anyone who has gone through the public schools is likely to have encountered the Iowa Tests of Basic Skills, Iowa Tests of Educational Development, Tests of Achievement and Proficiency, the Stanford Achievement Tests, the Comprehensive Tests of Basic Skills, or the Metropolitan Achievement Tests, all commercial standardized tests used in grades K-12.

What is standardized about standardized tests? The short answer is, almost everything. The format of all the questions for all students is the same. This format is usually, but not always, the multiple-choice format. All the questions for all students are the same. (Well into the 20th century, examinations were often oral, with each student getting different questions.) The instructions to all students are the same. The time permitted for all students to complete the test is the same. Standardized tests are often contrasted with "teacher made" tests, but even these share many of the standardized characteristics listed above.

Standardized tests are most often administered to groups of students, but some, such as IQ tests, can be given to individuals. In such cases, those who administer the tests have themselves been highly "standardized." That is, while they have some flexibility in the sequencing of questions and in applying the criteria for a correct answer, they must undergo extensive training to become standardized in the way the tests are given and scored. A child should not get an IQ of 100 from one test administrator and 130 from another.

About the only thing in the arena of standardized testing that is not standardized is the test-taker.

How Standardized Tests Are Constructed

The methods used to construct standardized tests are themselves highly standardized, especially the methods for constructing the most commonly used tests: commercially published achievement tests used by schools. To construct these tests, test publishers cull the most common textbooks and curriculum materials and try to develop questions that reflect the content of these materials. The questions are rated by curriculum specialists for what is termed "content validity," which is simply a rating by the experts as to whether they think the test actually measures what it claims.

The questions are then tried out on groups of people — standardized — to see if the questions "behave properly." Proper behavior in a test question is defined by means of statistical procedures.

We can see here one potential problem deriving from the way tests are constructed by test publishers, most

27

of which are subsidiaries of large publishing conglomerates: the process of construction can limit educational innovation. A test publisher naturally wants to sell a test to the widest possible market in order to obtain the largest possible profit. To have such an appeal, tests have to be oriented toward that which is common among schools, not that which is unique. This can inhibit educators from trying innovations that might depart from the common denominator of material that the test covers. If schools are evaluated on the basis of test scores, education reformers will be loath to make innovations that might not be reflected in these scores.

Now many people would hold that it is a good thing to learn about art or to play a musical instrument or to speak, read, and write a foreign language. But acquiring these skills will do nothing to improve students' performance on, say, the Iowa Tests of Basic Skills (ITBS). The vocabularies of art and music are too specialized to be a part of the reading or vocabulary sections of the ITBS. Specialized words in the lower grades do not lend themselves to the kind of items that have the statistical properties test-makers are looking for (more about just what they're looking for later). Similarly, learning Spanish might one day help a student to see the Latin root of an otherwise mysterious word on the SAT, but it will not do anything for the ITBS.

Norm-Referenced Tests

A norm-referenced test (NRT) is a standardized test with norms. And the norm is a rank, the 50th percentile. It is the rank assigned to the "median" score, and the

median is one kind of average. It is the median score of some group. For nationally used tests, the most common norm is a national norm constructed by testing children all over the country. There usually are norms for urban and suburban schools and, for some tests, private schools. In the case of the national norm, it is also the score that test-makers call "at grade level." By definition, then, half of all test-takers score at or above the 50th percentile, and half score below it. Half score above grade level and half score below it.

This property of norm-referenced tests bothers some people because, by definition, half of all test-takers will always be "below average." A system that labels half of our children as below average disturbs some people. It sometimes also leads to confusion — some people, politicians usually, can be heard decrying the fact that half of the students scored below average. The cries happen more frequently when the phrase "grade level" is used because it would seem intuitively obvious that everyone in, say, the seventh grade, should be at or above grade level. People who utter such cries do not realize that this outcome was guaranteed in advance by the way the test was constructed.

A norm-referenced test gives scores in relation to the norm, the 50th percentile, hence its name. If one of your students receives a report that says he scored at the 75th percentile, you know that he scored better than 75% of the students in the nation who took the test and that 25% of the students scored better than he did. You do not know from a percentile whether your student is doing well or poorly or average in any absolute sense.

Establishing the Norm. To determine the norm, test publishers first try out their questions on students and choose the questions that behave properly. By and large, this means choosing questions that 50% of the students miss. Some questions will be easier, and some will be harder. But rarely do tests include questions that 90% of the students get right or that 90% get wrong.

The choice of items that 50% of the students fail is an artifact of the history of testing in this country. If you want to make differential predictions, you have to arrange it so different people get different scores. If you choose items that everyone misses or everyone gets right, then everyone gets the same score, and you can not make differential predictions. It turns out that if you choose items that, on average, 50% of the test-takers get right and 50% get wrong, you end up with a test that distributes scores in a normal, bell-shaped curve and maximizes the dispersion of the scores.

The very presence of "distracters" — the wrong answers presented in multiple-choice questions — bothers a number of test critics. The test publishers must trick some students into choosing a wrong answer. If they cannot do that, then the item will not "behave" properly. That is, it will not be missed by half of the students. Leaving aside whether trying to trick students into making mistakes is an appropriate activity for educators, let us note that this procedure can be a barrier to good test construction under some circumstances. For instance, when I was a teaching assistant for introductory psychology at Stanford University in the 1960s, the administration wanted me to grade "on the curve,"

allowing 15% A's, 35% B's, 35% C's, and 15% D's and F's. This distribution approximates a normal, or bell-shaped, curve. Now, at that time the average total SAT score of entering freshmen at Stanford was about 1225, and 22% of them had never seen anything lower than an A on their report cards — and this was in the days before people started worrying about "grade inflation." Yet, at the directive of the university administration, 50% of these bright young people were supposed to be handed C's, D's, and F's.

The students knew this, of course, and they studied hard to be in the upper half of the distribution. This made it even more difficult to assign grades in line with the administration's wishes. If I asked questions about the important material in the chapter, everyone would get the questions right, and I would not have any differentiated scores to use as a basis for grading.

The discussion in the previous paragraphs reveals something important about tests: Much of what determines whether a question will be on a test has nothing to do with the content of the test. Whether a question gets on a test has to do with technical, statistical concerns about how the item "behaves."

What if a community has high percentile ranks? It could mean that all the kids are smart or that their school system is good or both. It is also possible that they are just rich. Wealthy communities in this country spend more on their schools, and families in such communities can provide enrichment at home that is unavailable to people in poorer communities. Some districts that are not average in wealth thus complain that comparing

themselves to the national norm is not meaningful. To assist these communities, test publishers develop various kinds of "local norms."

For example, a suburb can compare itself to other suburbs if it so chooses. It seldom does so because the rank of the suburb in comparison to similarly advantaged communities will not look as good as a comparison to the nation as a whole. On the other hand, an impoverished city neighborhood will look better when compared to impoverished neighborhoods in other cities than to the nation as a whole.

And that illustrates a problem with "local norms" in education: They can obscure real problems. An inner-city school that adjusts its test reports to take account of local norms can say, "We're doing as well as expected, and we'd be doing just as well as the rest of the country if we didn't have all these poor kids in our buildings."

Domains. We speak of tests of reading, mathematics, history, and so forth, as if they were generic tests of these subjects. But they are not. They are short tests of specific skills. Test publishers admit that their tests cannot cover everything. How could they? They are only 25 to 40 items long. The publishers will contend that their tests sample a larger "domain." The items sample only a small part of the overall domain. The theory of domain sampling is largely nonsense because no one has ever specified what a curriculum "domain" is. All the tests really do is ask specific questions in specific ways.

Multiple-Choice Format. The multiple-choice format is a peculiar way of measuring something. This can be seen

most readily in connection with writing. For a number of years, when multiple-choice tests were virtually the only form used, "writing" achievement was assessed in terms of what might be more accurately called editing skill. For instance, students would look at a sentence and decide what was wrong with it by choosing one of the four or five "corrections" provided. Or the test might show a short paragraph with four or five parts of it underlined, and the student's task was to decide which underlined section contained an error (or if none of them did). As a consequence, children did not learn how to write. When educators later came to their senses and realized that writing can be taught and assessed only by having kids write, children's writing improved immensely.

There is another aspect of multiple-choice tests that merits comment here, and it was captured wonderfully by one T.C. Batty in a 1959 letter to the *Times* of London:

> Sir,
>
> Among the "odd one out" type of questions which my son had to answer for a school entrance examination was, "Which is the odd one out among cricket, football, billiards, and hockey?"
>
> I say billiards because it is the only one played indoors. A colleague says football because it is the only one in which the ball is not struck with an implement. A neighbor says cricket because in all the other games the object is to put the ball into a net; and my son, with the confidence of nine summers, plumps for hockey "because it is the only one that is a girls' game."
>
> Could any of your readers put me out of my misery by stating what is the correct answer, and further enlighten me by explaining how questions of this sort prove

anything, especially when the scholar has merely to underline the odd one out without giving a reason?

Perhaps there is a remarkable subtlety behind all this. Is the question designed to test what a child of 9 may or may not know about billiards — proficiency at which may still be regarded as the sign of a misspent youth?

Yours faithfully,
T. C. Batty

There is not, unfortunately, any kind of "remarkable subtlety" behind such a question.

What You Test Is What You Get. One final condition of testing that deserves comment here is the Law of WYTIWYG (pronounced, wittywig): What You Test Is What You Get. There are two aspects of this law. The first and most obvious is that teachers will spend more time on the topics covered by a test than they will on topics not covered by the test, particularly if test scores are emphasized by the school staff, the school board, the media, or local industry. Thus it is very important to use tests that reflect instructional priorities and that do not in and of themselves cause the curriculum to become narrowed by teaching to the test. Teaching to most school tests is a problem because it constitutes "cheating."

But that is not the real problem. A high school football coach teaches to the test all week long, and we don't call it cheating. Indeed, we would think him crazy if he did anything else. But in this instance, the "test" is a real-life experience to get ready for. Many achievement tests, though, are in no way "real life" experiences. A teacher can teach to a test, but it comes at the expense of not teaching other parts of the curriculum.

There is a second and more subtle aspect of testing related to the Law of WYTIWYG: How you test determines, in part, what you see. When you measure people by means of a multiple-choice test, some people will do well and others won't. When you measure people by means of a performance assessment, some people will do well and others won't. But they won't be all the same people. Athletes often return from their "tests" saying, "I learned something out there today." Students seldom walk away from a testing situation with the same sense of accomplishment.

Criterion-Referenced Tests

How a person stands relative to the norm or to other people — a normative score — is not the only possible kind of score. However, normative scores are overwhelmingly the easiest to obtain and, as a consequence, are the most common. However, it is possible to score performance in relation to a clearly specified set of behaviors and, in the early 1960s, this fact led to the development of another type of standard test referred to as criterion-referenced tests (CRTs). Most of these tests were not really criterion referenced, but that gets us ahead of our story.

What is a "clearly specified set of behaviors"? Such behaviors are hard to specify for the topics taught in school. So most examples of clearly specified behaviors were taken from other areas. The inventor of the phrase, "criterion-referenced testing," Robert Glaser of the University of Pittsburgh, said that we could imagine achievement as a continuum of specified behaviors from

35

zero performance to conspicuous excellence and place any given performance somewhere along that continuum. For example, if ice skating was chosen as the clearly specified set of behaviors, then we could imagine a zero point as "Can't stand alone on ice." Conspicuous excellence at the other end of the continuum might be "Completes triple axel with perfect landing." A triple axel is a specific set of behaviors that judges can agree on with near perfect uniformity. In between zero and the triple axel are intermediate levels of accomplishment. The standards for these accomplishments (the criteria of CRT) can be described and the performance of the skater evaluated in reference to them.

The world of education is somewhat more vague and complex, but the concept of CRTs generated a great deal of enthusiasm. The notion of CRTs, in the words of Cornell University psychometrician Jason Millman, "totally destroyed the monopoly of norm-referenced interpretations that was held in many quarters." In 1994, Millman reflected on how he and many others reacted to the concept of a CRT:

> Thirty years ago I was a young pup, full of ambition and optimism. I thought that if only educators could write good test specifications, explicitly stating what was and was not part of the content coverage, CRTs would be able to meet their promise. More than that, even, I believed CRTs could give quantitative interpretations such as: Billy can answer 65% of the questions contained in a given domain. But I was wrong.[8]

Millman's enthusiasm and subsequent disillusionment were common. What happened? Basically, the

testers found that they could not specify educational outcomes with the clarity with which they could specify the outcomes of ice skating. Ice skating is easy to observe, but what is in a kid's head is not. It is also difficult to infer from a test precisely what the student knows. More important, ice skating is a very limited range of behaviors. The goals of education are more general.

Although criterion-referenced tests became all the rage in the 1970s, they had a problem: They had no criteria. Or, more precisely, a "criterion" was imposed on the test through the act of setting a "cut score" for passing or failing. The use of cut scores has itself been controversial and problematic, particularly in the use of "minimum competency" or other tests to determine grade promotion or eligibility for graduation. Such tests were tests of "minimal" skills or "essential" skills and so gave the impression that all test-takers should attain perfect scores, otherwise they would fail in life. So what did it mean, then, to set a cut score at, say, 60% correct, a commonly used figure?

More troublesome was the mere notion of minimum competency. Should a student who scored 61% be permitted to graduate while a student who scored 59% was forced to repeat his entire senior year or receive something less than a diploma?

These kinds of issues concerning cut scores were never really resolved — they are, in fact, not resolvable by technical means, though taking measurement error into account is technically useful. The issues were resolved in practice by setting the cut scores high enough

to ensure that enough students initially failed to satisfy those who had called for the tests in the first place, but low enough that very few students would not pass by the time of their graduation.

38

Performance Tests

Beginning in the late 1980s, some people in the assessment field did more than just express frustrations over the limits of multiple-choice tests. They began experimenting with tests that actually required students to perform. These tests became known popularly as "authentic tests," though some objected to this name because it implied that other forms of testing were "inauthentic" or phony. The term authentic was used because the tests required students to solve authentic problems from the real world, not situations in which a student had to select one of the alternatives provided by the test-maker. These tests were a type of performance test; and performance test is probably a better generic term, so we will go with that.

Performance tests, for the most part, are not "standardized" in the ways that the tests we discussed in the preceding section are "standardized." When all students respond to the same writing "prompt" in a writing assessment, then some amount of standardization is involved, especially if they must all finish in a given amount of time and are given no opportunity later to edit and revise. But performance tests introduce a degree of idiosyncrasy into the assessment situation.

Performance tests are one kind of criterion-referenced test that directly measures the performance the assessor is interested in. For example, music competitions, auditions, portfolios, and athletic contests all directly measure the skills of interest.

Why are performance tests so seldom seen in schools? There are several reasons. In all the examples of performance testing listed, only small numbers of people are involved at any one time. Performance tests take a lot of time and money to administer and score. If the goal of an assessment can be reached by using the much faster, much cheaper multiple-choice tests, then there is little reason to spend the extra time, effort, and money on performance assessments. And recall that, in the history of testing in this country, the emphasis has been on making discriminations among people, not on determining how well they actually perform. Paper-and-pencil tests can spread people out on bell curves much faster and cheaper than performance measures can. Similarly, if the interest is in obtaining some idea of how well a school or, more likely, a school system, is functioning, the use of performance assessments would be horrendously expensive and time-consuming.

However, there are many aspects of education (and life) that do not lend themselves well to multiple-choice tests. The most obvious school-related area has already been mentioned, writing. One cannot measure writing skills by means of multiple-choice questions. More important, children cannot learn to write by practicing the editing skills that can be assessed through multiple-choice tests. Children must actually write to learn how

to write. This statement seems so obvious that one wonders why it was ignored for so long.

Here is one reason why: Traditional testers would point out that writing assessments using actual writing and "writing" assessments using multiple-choice questions were highly correlated, so why bother with the more expensive, time-consuming writing? This is an adequate response only in situations where one wishes to make discriminations among students. Since that is what traditional testing has been all about, the response is not surprising. But the correlation between the two types of tests tells you only that those who score high on the multiple-choice test tend to score high on the writing test. "Scoring high," however, tells you nothing about the quality of the performance. "Scoring high" tells you only about the performance of students in relation to each other — it is a normative statement. It could easily be that the writing of all students is awful but that those who score higher than their peers on one form of the test also score higher than their peers on the other form.

The statement, "children must write to learn how to write," illustrates another aspect of good testing: Good testing does not simply measure performance, it informs it. People can learn something from it. Not much can be learned from taking a multiple-choice test.

Schooling should be, in part, about teaching people to think. The ability to critically evaluate information is crucial to functioning in a democratic society. Some have argued that multiple-choice tests can test higher-order thinking skills. It is certainly true that such tests

can test higher-order thinking skills, but they rarely do. Usually multiple-choice tests that require thinking are found in a limited range of courses in graduate school. Indeed, most tests punish the thinking test-taker. Thinking takes time. And as a test-taker, the last thing you want while taking most tests is something that slows you down the way thinking does.

We can see the import of thinking and the difficulty of using multiple-choice tests by describing higher-order thinking. Lauren Resnick, a cognitive psychologist at the University of Pittsburgh, has listed some of the qualities of higher-order thinking. Higher-order thinking:

- Is nonalgorithmic. That is, the path of action is not fully specified in advance. This is analogous to having to structure a problem.
- Tends to be complex. The total path is not "visible" (mentally speaking) for any single vantage point.
- Often yields multiple solutions, each with costs and benefits, rather than unique solutions.
- Involves nuanced judgment and interpretation.
- Involves the application of multiple criteria, which sometimes conflict with one another.
- Often involves uncertainty. Not everything that bears on the task at hand is known.
- Involves self-regulation of the thinking process. We do not recognize higher-order thinking in an individual when someone else "calls the plays" at every step.
- Involves imposing meaning, finding structure in apparent disorder.

- Is effortful. There is considerable mental work involved in the kinds of elaboration and judgment required.[9]

A student who deploys higher-order thinking of this kind while taking the SAT is in trouble. Time will expire before the test has been completed.

Performance tests are not without their problems, aside from their costs in time and money. For instance, there are ambiguities in the use of portfolios. Portfolios are used in a number of schools and districts as the major portion of writing assessment. Decisions have to be made as to what goes into a portfolio: the student's best work or the student's typical work? The teacher's selection of "best work" or the student's? How many writing types? Students with an aptitude for narration might want to concentrate on stories, ignoring essays, technical reports, and poetry. Who grades the portfolio? Different teachers judge the same work differently. While this last has been considered largely in terms of teacher unreliability, it probably also reflects genuine philosophical differences. One solution to this problem has been to train teachers to judge certain aspects of writing in similar ways. But does this cause them to overlook other meaningful qualities of the writing?

If the work is done in a group, but grades are to be assigned to individuals, the question arises as to whose work it is. Some parents object to group projects, alleging that a few kids end up doing all the work. Other parents accept the practice as good preparation for real life.

43

Multiple-choice tests will no doubt continue to flourish. For accountability and differentiation among students, it is hard to see how performance tests can replace them. Parents, it seems to me, should be concerned that performance tests become part of their children's educational experiences. An education geared only to the skills that are tested by commercial achievement tests, the SAT, and similar tests will be schooling aimed at the traditional goals of mass education.

44

Interpreting Test Scores

We have talked about test scores mostly in terms of percentile ranks, but ranks are not the only way of reporting test scores. Other ways include grade-equivalents, normal-curve equivalents, standard scores, and stanines. All of these scales are mathematically related to the normal curve shown in Figure 2 on page 47. Let's take a brief look at each of these in turn.

Grade-Equivalents

Teachers and parents love grade-equivalents. They have such an intuitive appeal. If Suzy is in the third month of the third grade and gets a grade-equivalent of 3.3 on a test, the teacher can tell the parents that Suzy is "at grade level," and the parents can go home thinking that Suzy is where she should be for her age. Test-makers define "grade level" as the score of the average student in a particular grade. But, like a national norm, 50% of all children are by definition below "grade level." This kind of definition can lead to a lot of mischief if, as has happened, a newspaper reports that 30% of the members of a graduating class in a high school are not reading at grade level. People who are not aware of how

test publishers define grade-equivalent will assume, quite naturally, that all of the graduating seniors should be reading at grade level. And if they aren't, what on Earth is the high school doing giving them diplomas? But, to repeat, half of all students in the nation will be below grade level — by definition.

Further — and more common — mischief often occurs when a child in, say, the fourth grade brings home a test report declaring that she has a grade-equivalent in reading of 7. Why, the parents are likely to wonder, is my child not in seventh grade, at least for her reading, since she is reading at seventh-grade level? But she is not reading at seventh-grade level. The seventh-grade level for seventh-graders is the score that the average seventh-grader would score on seventh-grade reading material. When a fourth-grader gets a grade equivalent of 7 on a test, it represents what the average seventh-grader would score on fourth-grade reading material.

Still more trouble arises when people average grade-equivalents. They look like they can be averaged. After all, they're numbers, aren't they?

Grade-equivalents may look like any other numbers, but they are not. Scientists talk of scales as having different properties that determine what can be done with them mathematically. In ascending order of precision, these scales are nominal, ordinal, equal interval, and equal ratio. A scale must attain the status of "equal interval" before its numbers can be meaningfully averaged.

In an "equal interval" scale the distance between one number and the next is always the same. In such a scale, numbers can be averaged. Temperature, whether in de-

grees Celsius or Fahrenheit, is such a scale. Each degree represents the same amount of heat. Taking one temperature reading of 60° and another of 40° and calculating an average of 50° is reasonable and gives a meaningful result. But averaging grade-equivalents is more like averaging house numbers (an ordinal scale). They are not part of an equal interval scale. You know that 608 S. Elm Street is farther south than 604 S. Elm, but 606 S. Elm (the average), might be quite close to 608 and quite far away from 604.

Percentiles

We have already discussed percentiles in passing, but we should mention here that they are not an equal interval scale, either. In terms of distance along the normal curve, a gain from the 50th percentile to 60th is much smaller than a gain from 80th to 90th. This is shown in Figure 2.

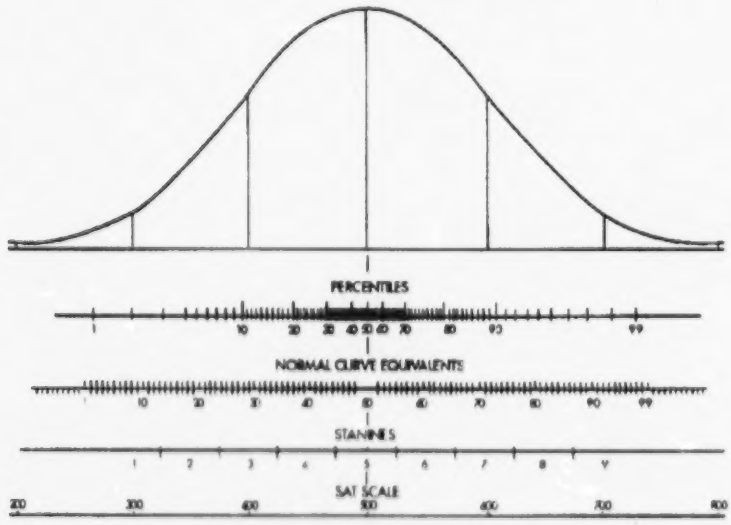


Figure 2. Scales related to the normal curve.

Normal-Curve Equivalents

The normal-curve equivalent (NCE) was developed to try and remedy the problems discussed above concerning grade-equivalents and percentile ranks. The NCE was an attempt to create an equal interval scale for testing. Instead, it just created incomprehensibility. If your student scores at the 60th percentile, you know he did better than 60% of the kids in the population tested. If he gets an NCE of 60, you could learn, by looking at the normal curve below, that he scored better than 68% of the other students. For a percentile rank, you don't need to look at the normal curve. But an increase of your child's NCE from 60 to 70 doesn't mean that he scored better than 78% of the students. It means he scored better than 83% of the other kids — something you could learn only by looking at the normal curve again. Well, you could memorize all of these numbers, but no one bothers when percentile ranks give them to you automatically.

Standard Scores

Standard scores are used everywhere, so it is a good idea to have some notion of where they come from and how they can be interpreted. In a way, standard scores are the most technical of the various types of scores, but they are also the most familiar. An IQ score is a standard score. So is the score on the SAT, ACT, NAEP, GRE, LSAT, GMAT, and so forth. Almost all commercial achievement tests come with some scale referred to as a "growth scale" or "developmental scale." These, too, are standard scores. Unlike percentile ranks, which tell

you only where your child stands in regard to other children, standard scores permit you to see how much progress your child has made between, say, third and fourth grades. School districts seldom make use of these kinds of standard scores.

All of the standard scores listed above begin life as a scale defined in terms of standard deviations and the normal curve. Suppose we have a bunch of test scores whose mean is 50 and whose standard deviation is 10. Then consider the selected set of scores: 20, 30, 40, 50, 60, 70, 80.

Suppose we take each of these scores, subtract the mean from it, and divide by the standard deviation. Here is what would happen to, say, a score of 80. If the mean is 50 and the standard deviation is 10, then a score of 80 is 30 points — three standard deviations — above the mean. And when we subtracted the mean and divided by the standard deviation, we got a new score in terms of standard deviation units: (80-50)/10 = +3. The whole set of sample scores would look like this:

(20-50)/10 = -3
(30-50)/10 = -2
(40-50)/10 = -1
(50-50)/10 = 0
(60-50)/10 = +1
(70 -50)/10 = +2
(80-50)/10 = +3

For all tests the standard scores, called z scores, will always range between -3 and +3. They are scores that measure a person's position on the normal curve. And

+3 means three standard deviations above the mean. Easy to remember. Much easier than juggling all those means and standard deviations for each test.

But we also said that IQ scores and SAT scores are standard scores, and these scores don't look anything like IQ or SAT scores. True, but we can easily get from what we have to those scores. Let's take each z score, multiply by 15, and add 100. A z score of 0 — the mean — becomes 0 x 15 + 100 = 100. A z score of +3 becomes 3 x 15 + 100 = 145, and so forth. The distribution of scores now looks like this: 55, 70, 85, 100, 115, 130, 145.

Now these are beginning to look a lot more like IQ scores: they have a mean of 100 and a standard deviation of 15. They could have a mean of 50 and a standard deviation of 5 — the choice of scale is completely arbitrary.

The conversion from raw scores to z scores does not change the relationship of the scores to one another in any way. If I know that a person got a z score of +1, I know he did better than 84% of the rest of the test-takers. How do I know this? Well, 50% of the scores fall below the mean (assuming a normal distribution here so that the mean, median, and mode are all the same). And 34% of all scores will fall between the mean and +1 standard deviation.

This is mostly something that needs to be memorized. Thirty-four percent of all scores will fall between the mean and +1 or -1 standard deviation. Another 14% will fall between +1 and +2 standard deviations and -1 and -2 standard deviations. So, in all normal curves, 96% of all scores will fall between -2 and +2 standard deviations.

50

Stanines

Stanine is short for "standard nine." Each stanine represents a certain percentage of the normal curve, as follows:

Stanine	1	2	3	4	5	6	7	8	9
Percentage	4	7	12	17	20	17	12	7	4

Thus the bottom 4% of the scores are the first stanine; the top 4%, the ninth.

Stanines are seldom used anymore, and people might be hiding something if they use stanines. It is common practice to consider the fourth, fifth, and sixth stanines, the three stanines in the middle, as "average." If you add up these three stanines, you get 54% of the scores. These can be reported as average. The seventh, eighth, and ninth stanines contain another 23% of the scores. So if you are an enterprising school public relations official, you can tell the media that 77% (54% + 23%) of your students are average or above average. Only the bottom 23% of the scores — those in the first, second, and third stanines — need be reported as "below average."

Conclusion

It is to be hoped that by the time you reach this point that you have a better idea of how tests are made, what they can and cannot do, and how to interpret the results from them. The technicalities of tests have not been covered because space is not available in this fastback. I sketch them out in the book from which this fastback is condensed, *Put to the Test: An Educator's and Consumer's Guide to Standardized Testing*. (Bloomington, Ind.: Phi Delta Kappa International, 1998.).

Anyone interested in pursuing them further might wish to peruse Lee Cronbach's *Essentials of Psychological Testing* (Harper & Row) or Anne Anastasi's *Psychological Testing* (Macmillan).

It is not likely that we will reduce our reliance on tests any time soon. Although President Clinton's proposal for a national test met, at best, a lukewarm reception in the statehouses around the country, it is not because the governors are anti-test. For the most part, they are pushing their own test development programs and simply think that such programs are better operated at the state rather than the national level. The best we can hope for is that you can use the knowledge gained from this book to keep tests in their proper perspective.

52

Notes

1. Roughly half, anyway. It depends on whether the college is using a "mean" or a "median" as its definition of "average." Highly selective colleges are more likely to use a mean; open enrollment colleges are more likely to use a median. All of this is explained in more detail in later sections.
2. Grant Wiggins, *Assessing Student Performance: Exploring the Purpose and Limits of Testing* (San Francisco: Jossey-Bass, 1993), p. 72.
3. F. Allan Hanson, *Testing Testing: Social Consequences of the Examined Life* (Berkeley: University of California Press, 1993), p. 179.
4. Desmond Hanerd, "The Failed Miracle," *Time*, 22 April 1996, p. 60.
5. "Reforming Japan's Schools," *Principal* (January 1998): 24-27.
6. Alexandra K. Wigdor and Wendell R. Garner, eds., *Ability Testing: Uses, Consequences, and Controversies, Part II* (Washington, D.C.: National Academy Press, 1982), p. 317.
7. Anne Anastasi, *Psychological Testing*, 6th ed. (New York: Macmillan, 1988), p. 414.
8. Jason Millman, "Criterion-Referenced Testing 30 Years Later: Promise Broken, Promise Kept," *Educational Measurement, Issues, and Practices* (Winter 1994): 19.
9. Lauren B. Resnick, *Education and Learning to Think* (Washington, D.C.: National Academy Press, 1987), p. 3.

Recent Books Published by the *Phi Delta Kappa Educational Foundation*

S.A.F.E. Handbook
Michael D. Rowland
Saddle-stitched paperback. $8 (PDK members, $6)
Special Offer: 10 copies for only $56

Planning for Disaster, Second Edition
Harmon A. Baldwin
Saddle-stitched paperback. $8 (PDK members, $6)
Special Offer: 10 copies for only $56

Inclusion: Policy and Practice
Thomas P. Lombardi, ed.
Trade paperback. $13 (PDK members, $9.75)

American Overseas Schools
Robert J. Simpson and Charles R. Duke, eds.
Trade paperback. $22 (PDK members, $16.50)

Getting a Job in Teaching
Joanne C. Wachter
Saddle-stitched paperback. $8 (PDK members, $6)

Use Order Form on Next Page
Or Phone 1-800-766-1156

A processing charge is added to all orders.
Prices are subject to change without notice.

Complete online catalog at http://www.pdkintl.org

Order Form

SHIP TO:	
STREET	
CITY/STATE OR PROVINCE/ZIP OR POSTAL CODE	
DAYTIME PHONE NUMBER	PDK MEMBER ROLL NUMBER

QUANTITY	*TITLE*	*PRICE*

ORDERS MUST INCLUDE PROCESSING CHARGE

Total Merchandise	*Processing Charge*
$3 to $25	$3
$25.01 to $100	$5
Over $100	5% of total

Special shipping available upon request. Prices subject to change without notice.

SUBTOTAL	
Indiana residents add 5% Sales Tax	
PROCESSING CHARGE	
TOTAL	

☐ Payment Enclosed (check payable to Phi Delta Kappa International)

Bill my ☐ VISA ☐ MasterCard ☐ American Express ☐ Discover

ACCT # _____ DATE _____

EXP DATE ___/___ SIGNATURE _____

Mail or fax your order to: Phi Delta Kappa International, P.O. Box 789, Bloomington, IN 47402-0789. USA
Fax: (812) 339-0018. Phone: (812) 339-1156

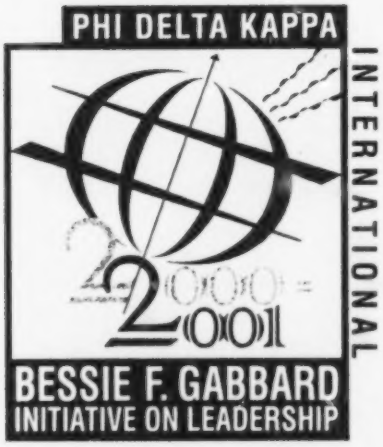
For fastest service, phone 1-800-766-1156 and use your credit card.

Phi Delta Kappa Fastbacks

This series, published each fall and spring, offers short treatments of a variety of topics in education. Each fastback is intended to be a focused, authoritative work on a subject of current interest to educators and other readers. Since the inception of the series in 1972, the fastbacks have proven valuable for individual and group professional development in schools and districts and as readings in undergraduate and graduate teacher preparation classes. More than 450 titles in the series have been published, and more than eight million copies have been disseminated worldwide.

For a current list of available fastbacks and other publications, please contact:

Phi Delta Kappa International
P.O. Box 789
Bloomington, IN 47402-0789 U.S.A.
1-800-766-1156
(812) 339-1156
http://www.pdkintl.org



Bessie F. Gabbard Initiative on Leadership

The Bessie F. Gabbard Initiative on Leadership in Education for the 21st Century, dubbed the 2000-2001 Celebration for short, reaffirms the central importance of the Phi Delta Kappa tenet of leadership. Bessie F. Gabbard, the "First Lady" of PDK and a member and longtime chair of the board of governors of the Phi Delta Kappa Educational Foundation, provided the impetus for this initiative, which will focus the energies of PDK members and staff during the two years of transition to the new millennium. During this 2000-2001 Celebration, special attention will be paid to leaders and leadership in education with a particular focus on PDK's traditional advocacy on behalf of the public schools.

ISBN 0-87367-659-9

Deaf students at the Alexander Graham Bell School in Cleveland, Ohio, enjoy reading in this early 20th-century class.

Courtesy of the Cleveland Public Library Photograph Collection.